Published: Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy. - Paris: European Language Resources Association, 2006. Pp. 471-474.

The Collection of Distributionally Idiosyncratic Items: A Multilingual Resource for Linguistic Research

Manfred Sailer* **and Beata Trawiński†**

*University of Göttingen
Department of English Studies
Käte-Hamburger-Weg 3
D-37073 Göttingen
manfred.sailer@phil.uni-goettingen.de

†University of Tübingen
Sonderforschungsbereich 441
Nauklerstraße 35
D-72074 Tübingen
trawinski@sfs.uni-tuebingen.de

Abstract

We present two collections of lexical items with idiosyncratic distribution. The collections document the behavior of German and English bound words (BW, such as English *headway*), i.e. words which can only occur in one expression (*make headway*). BWs are a problem for both general and idiomatic dictionaries since it is unclear whether they have an independent lexical status and to what extent the expressions in which they occur are typical idiomatic expressions. We propose a system which allows us to document the information about BWs from dictionaries and linguistic literature, together with corpus data and example queries for major text corpora. We present our data structure and point to other phraseologically oriented collections. We will also show differences between the German and the English collection.

1. Introduction

The *Collection of Distributionally Idiosyncratic Items* (CoDII) is a linguistic resource on lexical items which have highly idiosyncratic occurrence patterns. CoDII consists of two parts: bound words of German (CoDII-BW.de, accessible online since 2004, and a corresponding collection for English (CoDII-BW.en) which will be accessible starting in May 2006.[1]

Bound words (BW) are words such as *headway* which can only occur as part of a fixed expression (here: *make headway*). The repertoire of German BWs is well documented in literature on idioms. Dobrovol'skij (1988) contains the most exhaustive list of BWs for German, English and Dutch. Dobrovol'skij (1988, 1989) and Dobrovol'skij/Piirainen (1994a,b) provide criteria for classifying BWs and the expressions within which they occur. An important emphasis of these publications is the difference between bound and free words. Dobrovol'skij and Piirainen estimate the number of potential BWs for German as 600, out of which they classify 180 as belonging to the common vocabulary of native speakers. At present, we have included 450 potential BWs in our collection.

A thorough documentation of BWs is important because distributionally idiosyncratic items pose a serious challenge for both theoretical linguistics and large-coverage computational applications. On the one hand these items behave like "free words" syntactically and semantically, on the other hand, they show strong usage constraints, usually absent from implemented grammars. CoDII aims to present the linguistic knowledge available for these items. This is combined with corpus data which can be used to evaluate linguistic theories and to serve as a test basis for practical systems.

2. Design and Realization

Based on Dobrovol'skij (1988), all potential BWs were included in CoDII-BW.de. For each item linguistic documentation and examples from dictionaries and corpora are included. The linguistic documentation consists of the following four information blocks:

1. General Information: A particular BW is identified, and its English translation is given. We indicate the expression in which the BW occurs and provide a set of possible paraphrases of this expression.
2. Classification: We specify the classification of a given BW according to Dobrovol'skij (1988, 1989), Dobrovol'skij/Piirainen (1994b), and a classification based on Nunberg et al. (1994).
3. Syntactic Information: We provide the syntactic category of the BW and the syntactic structure of the expression in which it occurs. We list possible syntactic variations (passivization, pronominalization, modification, occurrence in raising constructions, etc.). For each context, examples from corpora, Internet or literature are included in the data section of the collection.
4. Sample Queries: We specify optimized sample queries for various publicly available corpora. Again, pointers to examples of selected query results are included.

We are grateful to Garrett Hubing for proofreading this paper.

[1] Both collections can be entered at `http://www.sfb441.uni-tuebingen.de/a5/codii`.

This information is internally coded in XML. The DTD has been specified in the following way:
The element `codii` is the document root. Different collections are identified by attributes `type` (to specify the collection type) and `xml:lang` (for the language of the data). The content model of the element `codii` consists of two elements: `dii-list`, whose content is a list of distributionally idiosyncratic items, and `dii-examples`, which contains a list of examples.
The content model of the element `dii-list` consists of a list of `dii-entry` elements. The content model of a `dii-entry` element consists of a set of elements which encode the four information blocks mentioned above: The content of a `dii` element gives the lexical item itself in the original language and, where possible, an English translation. A `dii-expression` element gives the expression in which the BW occurs. We collect the documentation of the item in a `dii-classification` element. Here, we indicate which information different dictionaries give (and if they list the particular item at all), and how the given item was classified in linguistic literature. Information about the syntax of the item and its expression are given in a `dii-syntax` element. Finally, a `dii-queries` element contains example queries for the given BW. For each of these pieces of information, we include pointers to relevant examples from the `dii-examples` element.
Figure 1 presents a fragment of the element `dii-entry` describing the German BW *Zampano* 'golden boy'.
The content model of the element `dii-examples` consists of a list of `example` elements linked with appropriate entries by means of the attributes `dii` and `id`. Figure 2 shows the element `example` describing a corpus example for the BW *Zampano* 'golden boy'.
For syntactic descriptions of BWs and their associated expressions we used the Stuttgart-Tübingen Tagset (STTS) (`http://www.sfs.uni-tuebingen.de/Elwis/stts/stts.html`). The examples in CoDII currently stem from corpora of the Institut für Deutsche Sprache (IDS) (`http://www.ids-mannheim.de/cosmas2/`), the corpus of the Digitales Wörterbuch der Deutschen Sprache (DWDS, `http://www.dwds.de/`), TIGERSearch, a search engine for retrieving information from a database of graph structures (`http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/`), and the Internet via Google.
Figure 3 shows the browser display of one of the 450 entries of CoDII-BW.de. The user interface displays all the linguistic information for the particular BW. Comments, information about the classification systems and the relevant examples can be obtained by clicking the links in this display. Not all 450 entries in CoDII-BW.de are equally exhaustively filled in, but the flexible XML encoding makes it possible to add more information and to update the collection using an XSL transformation.

```
<dii-entry id="zampano">
    <dii>
      <ol>Zampano</ol>
      <en>golden boy</en>
    </dii>
    <dii-expression>
      <ol>der gro&#223;e Zampano</ol>
      <en>the big doer</en>
    </dii-expression>
    <dii-classification>
      <dii-class class="dekompo" type="A5">
         <bibliography bib-item="A5"/>
      </dii-class>
    </dii-classification>
    <dii-syntax hits="zampano-Bsp
              zampano-apposition" cat="NE">
       <dii-expression-syntax cat="NP">
         der/ART gro&#223;e/ADJA Zampano/NE
       </dii-expression-syntax>
       <variation kind="OPEN"
                hits="zampano-ecclestone">
          <comment status="external">
             Spitzname von Formel-1-Manager
             Bernie Ecclestone
          </comment>
       </variation>
    </dii-syntax>
    <dii-queries>
      <query type="cosmasII">
          <query-text><![CDATA[Zampano]]>
          </query-text>
       </query>
    </dii-queries>
</dii-entry>
```

Figure 1: The CoDII-XML-encoding of *Zampano*

```
<example dii="zampano" id="zampano-Bsp">
  <source corpus="cosmasII">
    R97/APR.32703 Frankfurter Rundschau,
    29.04.1997, S. 15, Ressort: WIRTSCHAFT;
    F&#252;r eine lohnende &#220;bernahme
    sind einige H&#252;rden zu nehmen
  </source>
  <ol>
    "Ich glaube nicht, da&#224; da Manna vom
     Himmel f&#228;llt und der gro&#224;e
     Zampano f&#252;r diverse neue Stellen
     sorgt", meint der Betriebsratschef der
     Vegesacker Werft, Wolfgang Dettmer.
  </ol>
</example>
```

Figure 2: The CoDII-XML-description of a corpus example for *Zampano*

3. CoDII for English

The German part of CoDII has been complemented with an English version. This collection is based on the approximately 100 English BWs in Dobrovol'skij (1988). The underlying XML encoding of CoDII-BW.de proved suitable for an expansion to English. The differences between the collections reflect different degrees of linguistic documentation for the two languages:
We used the syntactic annotation scheme from the *Syn-*

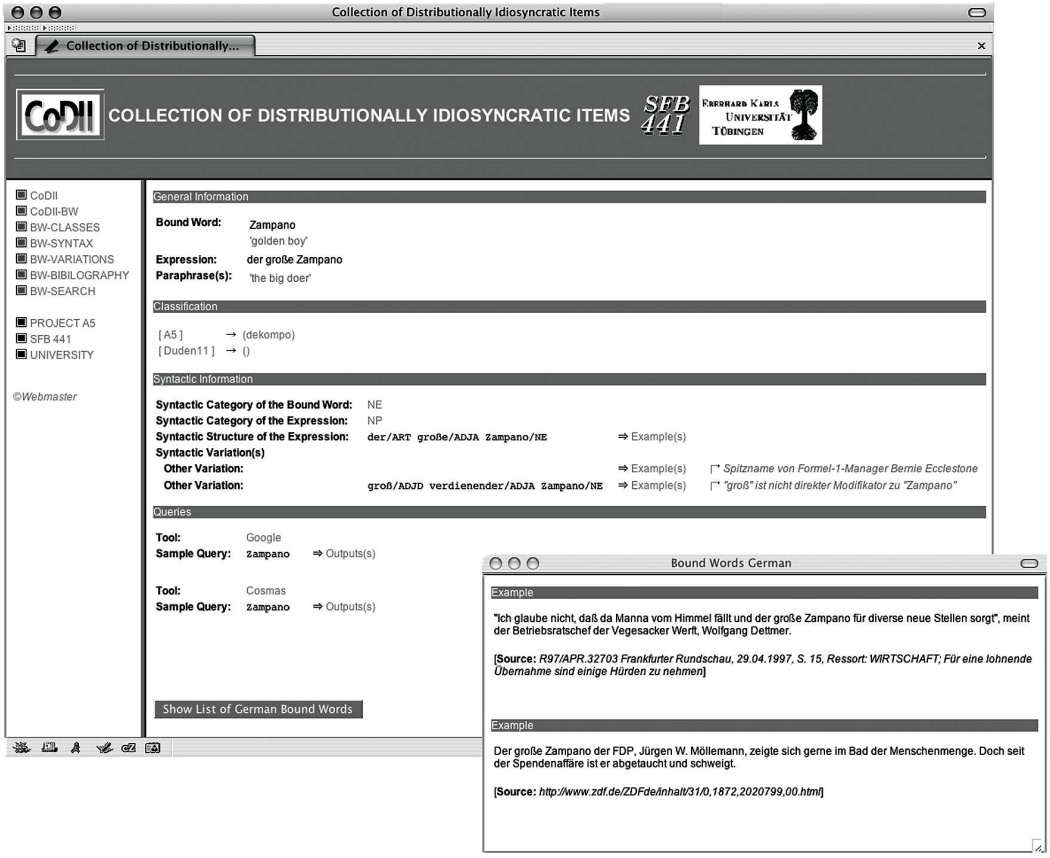

Figure 3: Browser display for the entry *Zampano*

tactically Annotated Idiom Database (SAID, Kuiper et al. (2003)) since this collection seemed to be the best reference point for a further study of the syntactic behavior of expressions with BWs. Due to the small amount of attention that BWs have received within English phraseological literature, we focused on the information given in idiom dictionaries and regular dictionaries. The only classification directly available for English BWs was that of Dobrovol'skij (1988). Additionally, we applied the classification of Dobrovol'skij/Piirainen (1994a), which is a classification of expressions rather than of individual BWs. The examples in the English collection stem mainly from dictionaries, the Internet and from the British National Corpus (via the SARA software package; `http://www.natcorp.ox.ac.uk/sara/`).

The German collection was developed in a project involving a small number of researchers who were at the same time developing the architecture of the collection. In contrast to this, the data for the English collection has mainly been compiled by a group of 25 advanced undergraduate students of English at the University of Göttingen. This was possible because the clear XML data structure served as a guideline. Some students investigated aspects of individual items, others integrated information from different dictionaries. The structure of CoDII allowed for a direct integration of the data and the observations from these studies. It is worth mentioning that even though the majority of the students had no previous experience with corpus research nor with XML, they were able to contribute independently to the collection after only a 4 hours tutorial. The document verification tools of XML editors were extremely helpful, here. (The majority of the students used the free XML editor cooktop, `http://www.xmlcooktop.de`)

4. Similar Collections

CoDII aims to serve as a database for linguistic research: Existing linguistic analyses have been compiled and can be extended, and CoDII is becoming a multilingual resource. Several other projects have constructed resources for idiomatic expressions. These projects differ from CoDII by the corpora used, the kind of data and the applied methods. *Usuelle Wortverbindungen* (Conventionalized Word Combinations[2]) of the IDS (Steyer, 2004) starts from statistically highly frequent words which undergo a co-occurrence analysis. This analysis serves as the basis for a linguistic and lexicographic description of the typical usage patterns of a word. In contrast to this collection, CoDII is based on linguistic intuitions and theoretical considerations. In part, this is due to the low frequency of a number of BWs. Another important difference is that IDS project only uses the corpora of the IDS — to have full control over the frequency data. For CoDII we try to collect as much information as possible about a given item. For this reason we want

[2] `http://www.ids-mannheim.de/lexik/UsuelleWortverbindungen`

to include data from different sources and retrieval strategies for different corpora.

Kollokationen im Wörterbuch (Collocations in the Lexicon[3]) of the Berlin-Brandenburgische Akademie der Wissenschaft (Fellbaum et al., ta) is based on the DWDS corpus. Similar to CoDII, the project starts with idioms from phraseological literature, but focuses exclusively on German VP idioms.

For English, the *Syntactically Annotated Idioms Database* (SAID, Kuiper et al. (2003)) encodes the syntactic structure of a huge number of idioms. However, it does not include any other information about the expressions. The SAID can be used to investigate structural generalizations about idioms. For this reason its encoding was also used for representing syntactic structures in the English CoDII-BW.

Villavicencio et al. (2004) present an interface for an interactive multilingual resource which is primarily designed to encode translation equivalents.[4] The interface has a clear and multi-lingual architecture. It allows external users to contribute to the collection. Unfortunately, there does not seem to be an extensive input from outside users. For this reason we decided not to have an interactive mode for inserting information from outside users directly into the database. However, we provide contact information and try to include the input we receive.

5. Outlook

The data structure design of CoDII makes it possible to add further classifications, corpora and search tools, as well as further collections of distributionally idiosyncratic items. Our experience with the creation of the English collection shows that such information can be added as part of small focused research projects by students after a short training period.

An extension to more languages is equally possible. A natural candidate for another language would be Dutch, for which Dobrovol'skij (1988) lists a large number of BWs. Furthermore, Feyaerts (1994) presents a detailed investigation and classification of Dutch BWs.

It is also planned to extend the collection to other types of distributionally idiosyncratic items. In particular, we intend to include a documentation of the use of polarity items (such as English *any*, or *lift a finger*), i.e., items which require a negative context.[5] These items represent an even greater challenge for traditional lexicography and computational applications because the occurrence requirements are not as local as in the case of BWs and because the obligatory collocators are not simple lexemes but abstract grammatical and semantic categories (van der Wouden, 1997).

On the technical side, CoDII will be converted into a database to allow for a dynamic and more flexible access to the data. This database will then be integrated into the TUSNELDA collection of the Collaborative Research Center 441.

[3] `http://www.bbaw.de/bbaw.Forschung/Forschungsprojekte/kollokationen/en/Startseite`

[4] `http://lingo.stanford.edu/cgi-bin/annotate.mli.cgi`

[5] Gianina Iordăchioaia has been working on a prototype for Romanian polarity items in the context of the project A5 of the Collaborative Research Center 441, University of Tübingen.

References

Dobrovol'skij, D. (1988). *Phraseologie als Objekt der Universallinguistik*. Leipzig, Enzyklopädie.

Dobrovol'skij, D. (1989). Formal gebundene phraseologische Konstituenten: Klassifikationsgrundlagen und theoretische Analyse. In W. Fleischer et al. (Eds.), *Beiträge zur Erforschung der deutschen Sprache*, Volume 9, pp. 57–78. Leipzig, Bibliographisches Institut.

Dobrovol'skij, D., Piirainen, E. (1994a). PGF: Auf dem Präsentierteller oder auf dem Abstellgleis? *Zeitschrift für Germanistik* (NF 4), 65–77.

Dobrovol'skij, D., Piirainen, E. (1994b). Sprachliche Unikalia im Deutschen: Zum Phänomen phraseologisch gebundener Formative. *Folia Linguistica 27*(3–4), 449–473.

Fellbaum, Ch., Kramer, U., Neumann, G. (t.a.). Corpusbasierte lexikographische Erfassung und linguistische Analyse deutscher Idiome. In A. Häcki-Buhofer (Ed.), *EUROPHRAS 2004*.

Feyaerts, K. (1994). Zur lexikalisch-semantischen Komplexität der Phraseologismen mit phraseologisch gebundenen Formativen. In Chlosta et al. (Ed.), *Sprachbilder zwischen Theorie und Praxis*, pp. 133–162. Bochum.

Kuiper, K., McCann, H., Quinn, H., Aitchison, Th., Veer, K. van der (2003). Syntactically Annotated Idiom Database (SAID) v.1. Documentation to a LDC resource.

Nunberg, G., Sag, I. A., Wasow, Th. (1994). Idioms. *Language 70*, 491–538.

Steyer, K. (2004). Kookkurrenz. Korpusmethodik, linguistisches Modell, lexikographische Perspektiven. In K. Steyer (Ed.), *Wortverbindungen — mehr oder weniger fest*, pp. 87–116. Berlin/New York, de Gruyter.

Villavicencio, A., Baldwin, T., Waldron, B. (2004). A Multilingual Database of Idioms. *Proceedings of LREC 2004*, pp. 1127–30.

Wouden, Ton van der (1997). *Negative Contexts. Collocation, Polarity and Multiple Negation*. Routledge, London.